PB99-149585

DOT/FAA/AR-99/40

Office of Aviation Research
Washington, DC 20591

# Validating the Computer-Based Training Process for Aviation Security Screeners

J. L. Fobes, Ph.D.
Eric C. Neiderman, Ph.D.

Aviation Security Human Factors Program,
AAR-510
William J. Hughes Technical Center
Atlantic City International Airport, NJ 08405

March 1999

This report is approved for public release and is on file at the William J. Hughes Technical Center, Aviation Security Research and Development Library, Atlantic City International Airport, NJ 08405.

This document is also available to the U.S. public through the National Technical Information Service (NTIS), Springfield, VA 22161

U.S. Department of Transportation
Federal Aviation Administration

REPRODUCED BY: NTIS
U.S. Department of Commerce
National Technical Information Service
Springfield, Virginia 22161

## NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturer's names appear herein solely because the information is essential to the objective of this report.

**Technical Report Documentation Page**

| 1. Report No.<br>DOT/FAA/AR-99/40 | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| 4. Title and Subtitle<br><br>Validating The Computer-Based Training Process for Aviation Security Screeners | | 5. Report Date<br>March 1999<br><br>6. Performing Organization Code AAR-510 |
| 7. Author(s)<br><br>J. L. Fobes, Ph.D. and Eric C. Neiderman, Ph.D. | | 8. Performing Organization Report No. |
| 9. Performing Organization Name and Address<br><br>U.S. Department of Transportation, Federal Aviation Administration<br>William J. Hughes Technical Center<br>Atlantic City International Airport, NJ 08405 | | 10. Work Unit No. (TRAIS)<br><br>11. Contract or Grant No. |
| 12. Sponsoring Agency Name and Address<br><br>U.S. Department of Transportation, Federal Aviation Administration<br>Associate Administrator of Civil Aviation Security, ACS-1<br>800 Independence Ave., S.W.<br>Washington, DC 20590 | | 13. Type of Report and Period Covered<br>Draft<br><br>14. Sponsoring Agency Code<br>ACS-1 |

15. Supplementary Notes: Draft prepared by:
William Maguire, Ph.D. & Joshua Rubinstein Ph.D.
Federal Data Corporation
Science and Engineering Division
500 Scarborough Drive
Egg Harbor Township, NJ 08234

16. Abstract
Test performance data from Safe Passage's Computer-Based Training (CBT) system for aviation security screeners were used to evaluate multiple choice content and X-ray image interpretation mastery test items to assess initial screener training. Data from 8,366 CBT tests were used from 691 screeners at 3 different sites. Test questions were evaluated for readability, item-to-test correlations, error rates, quality of response options, relevance to job requirements, training content reliability, adverse impact due to race and gender, and validity. Implications of these results for the design of the Screener Readiness Test and the evaluation of new CBT systems is discussed.

| 17. Key Words<br>Aviation Security<br>Computer-Based Training (CBT) | 18. Distribution Statement<br>This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161 | | |
|---|---|---|---|
| 19. Security Classif. (of this report)<br>Unclassified | 20. Security Classif. (of this page)<br>Unclassified | 21. No. of Pages<br>23 | 22. Price |

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PROTECTED UNDER INTERNATIONAL COPYRIGHT
ALL RIGHTS RESERVED.
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE

## CONTENTS

## TABLES

## ACRONYMS

| | |
|---|---|
| ATL | Atlanta Hartsfield International Airport |
| CBT | Computer-Based Training |
| DIF | Differential Item Functioning |
| FAA | Federal Aviation Administration |
| FKR | Flesch-Kincaid Readability |
| SPI | Safe Passage International |
| SRT | Screener Readiness Test |
| TIP | Threat Image Projection |

# 1. INTRODUCTION

## 1.1 Background

The effectiveness of the national civil aviation security system is highly dependent upon the performance of people employed as checkpoint screeners. The training of these individuals is critical to their performance on the job. The Federal Aviation Administration (FAA) is accordingly interested in enhancing screener training and further improving their readiness for the job.

Federal Aviation Regulations (FAR § 108.17: Use of X-ray systems) require a program for initial and recurrent training of operators of X-ray systems. This includes training in the efficient use of X-ray systems and the identification of weapons and other dangerous articles. Section XIII of the Air Carrier Standard Security Program presents the standards for training and testing of persons performing screening and security functions.

In April 1997, the FAA approved the use of a Computer-Based Training (CBT) system for initial screener training. This system was developed by Safe Passage International (SPI) and consists of instruction modules for security checkpoint procedures. The screener trainee's performance is evaluated with a short test following each module, a final 50 item multiple choice test, and a 50 item image interpretation test consisting of X-ray images of bags containing or not containing threats. A thorough quantitative and qualitative evaluation of SPI's tests is possible because of the large volume of data that has been collected from CBT installed at 19 major US airports.

Other aviation screener CBT systems are being developed and are likely to also have mastery testing similar to that used in the SPI training (Neiderman & Fobes, 1998). The present evaluation will provide useful information for training developers of these new CBT systems for aviation screeners. The FAA's Aviation Security Human Factors Program is also developing a Screener Readiness Test (SRT) consisting of multiple choice questions and an X-ray interpretation test. This test can serve as the exit criterion to complete initial screener training and its development will also benefit from the findings presented here.

## 1.2 Purpose

This document describes the analytical process necessary to validate the test items in CBT systems for aviation screeners. Data from SPI's system are used here to illustrate how to address issues such as the following. Item evaluation - what factors affect error rates, item discrimination, and overall item quality? Reliability - do the test items show good reliability with this population and how should the reliability of a test be assessed in which questions are randomly sampled from a pool of questions? Fairness - does this test have adverse impact on specific racial groups and, if so, what strategies can be employed to evaluate racial bias and minimize adverse impact with these types of questions? Validity - does performance on this test predict on-the-job performance?

## 2. METHOD

### 2.1 Sample Characteristics

A sample of 8,366 CBT tests (unit, content, and image tests) from 691 screeners at 3 sites [Atlanta Hartsfield International Airport (ATL), Dallas/Fort Worth International Airport, and Seattle-Tacoma Airport] was used for the analysis of test characteristics. For 208 of these screeners, demographic information was available because they provided such information before taking the CBT. This sample was used to analyze adverse impact of the tests. There were 76 screeners, mostly from ATL, for which both CBT and a minimal amount of TIP data were available. A sample of 213 image tests was used for X-ray image item analysis.

### 2.2 Testing Process

A screener in training must pass each SPI test for the six instructional modules in sequential order (being prohibited from working on the next unit until the previous unit is passed). The 6 SPI unit tests each consisted of 10 multiple choice items sampled from a larger item pool. The unit tests could be repeated many times until they were passed. When all the units have been successfully completed, the screener took a 50 item multiple choice content test. Like the unit tests, screeners were permitted to take the seventh test on overall content repeatedly until they passed it. Unit tests and the content test were drawn from the same pool of 179 multiple choice items.

After screeners passed the final content test, they took an X-ray image, threat detection performance test. The image test consisted of 50 items drawn from a large image library of images classified into five separate categories. Innocent bags, guns, knives, FAA test items, and shields were presented in ratios of 30:3:3:7:7, respectively, Like the other tests, screeners could take the image test repeatedly if they did not pass. Unlike the other tests. However, screeners were not likely to see items repeated, when taking the test multiple times, because of the large size of the image library. After passing this test, they have successfully completed training.

With this method, students who took tests one through seven at least once may have been given the same questions more than once. Because students were able to retake each unit multiple times until they have passed that unit, those who passed each unit on the first attempt will have taken each question no more than two times (on the individual instructional unit test then again on the overall content test). Students who retook a unit multiple times, on the other hand, could have been exposed to the same question many times. Averages calculated using all unit test scores would be more influenced by those students who perform below average. This is because the less able students, who retook each unit several times before passing, had more of their test scores contributing to that average than did the students who passed the unit tests on the first attempt. For this reason, error rates and pass/fail rates were determined using the outcome only from the first time each test question was given to each student.

## 2.3 Data Structure

The CBT database contained the following information for each student: screener identification, whether they successfully completed the CBT, scores on each unit test, scores on final image and content tests, items included in each test and whether answered correctly or incorrectly. From this raw information, the following were calculated: the mean score and distribution of scores for each unit, item difficulty measured by the error rate of each item, item discrimination measures including item to total correlations, and the difference in item performance between candidates who were successful and those who were unsuccessful in completing training.

## 2.4 Item Readability

Content quality and readability of the test questions were assessed. Item readability was measured using the Flesch-Kincaid Readability (FKR) Index, a US Government Department of Defense standard, which assesses readability of text by calculating a grade level. A readability grade level was calculated for each test question using the formula,

$$\text{Grade level} = (L \times 0.39) + (N \times 11.8) - 15.59,$$

where L = the average sentence length (number of words / number or sentences) and N = the average number of syllables per word (number of syllables / number of words). (The calculation of readability provided with Microsoft Word 97 was not used because the formula it uses does not provide accurate Flesch-Kincaid readability scores with multiple choice test formats.)

Additional measures of item quality taken include grammaticality, semantic content, and wording assessed and compared to the item's error rate. Each question was categorized as to whether a) it was grammatically correct, b) its semantic content made sense, and c) the wording of the item was confusing.

## 2.5 Test Reliability

Test-retest reliability and internal validity were calculated for the tests. The structure of these tests, however, made these calculations difficult. When screeners took tests more than once, this normally was because they failed a test. They then reviewed the instructional materials before taking it again. Such uncontrolled effects of learning between replications of the tests confound reliability of the test with learning of new material.

Normally, either coefficient alpha or the Kuder-Richardson formula is used to measure the internal validity of tests. Because all of the unit and final tests consist of a random sample of items from a larger item pool, the item content of the test varies each time it is presented. For this reason, it is not possible to use the standard formulas or the standard statistical software to compute the internal validity of the tests. The approach to this problem instead was to determine average inter-item correlations. Using these values and the Spearman-Brown formula, it is possible to estimate internal validity.

It is possible to estimate the test-retest reliability of the tests, however, with the same pool of items for both the unit and final exams. If trainees passed the unit test, they were not subjected to further instruction in the specific content of that unit. Therefore, test-retest reliability can be determined for individual items by restricting the sample to those cases of an item's last appearance on the unit test and final tests. Test-retest reliability of the unit and final tests can be estimated from these item correlations by using the average test–retest correlation for an item and extrapolating to the whole test using the Spearman-Brown formula.

## 2.6 Test Fairness

*The Uniform Federal Guidelines on Employee Selection Procedures* (1978) require efforts be made to ensure the fairness of any test to be used for employee selection. By evaluating the fairness of the CBT content and image tests, potential fairness issues are anticipated which may arise when using similar test and item structures.

At some sites, screeners took a set of tests that are being investigated as part of a potential screening battery before beginning CBT (Fobes & Neiderman, 1998). As part of that testing process, they completed a personal information form which provides information about the demographic variables of: gender, ethnic background (Asian, black, Hispanic, other, and white), native or nonnative English speaker, level of education, and age.

The effect of demographic variables on CBT scores was examined to determine whether demographics of race or gender show an adverse impact from the CBT's questions. Two basic measures of CBT performance were examined. One was the CBT success rate, the number of individuals who successfully completed the CBT as a proportion of the number of individuals who started it. The other variable was the score on the tests the first time they were taken.

Sufficient data were available to examine the Differential Item Functioning (DIF) analysis of racial/gender effects on item performance where ability is explicitly equated between groups. The best known approach to DIF is the application of item response theory (Hulin, Drasgow, & Parsons, 1993), but this approach demands very large samples. A more parsimonious analysis can be carried out using logistic regression (Camilli & Shepard, 1994). The approach described here was to construct stepwise logistic regression analyses where the dependent variable to be predicted was item performance (pass/fail) and the predictor variables were ability (using the test score as a measure of overall ability) and a categorical race variable (non-white, white). Significant contributions of the race variable to logistic regression beyond the contribution of ability (which is generally a significant predictor) were examined. Those questions where the race category contributed significantly to the odds ratio in the regression, were generally questions where racial differences in item difficulty are found across ability levels. An examination of the cross tables indicated whether the racial effect is biased against the non-white group.

## 2.7 Test Validity

At some sites, screeners who had successfully completed training and are on the job are using X-ray systems equipped with Threat Image Projection (TIP). The TIP systems present the images of threats superimposed on the real bag images on the X-ray screen. Screeners press a 'threat' button when they suspect a threat and, in this way, threat detection performance data become available for these screeners. Four measures of individual screener performance were used. These were the a) hit rate - the proportion of TIP images that were correctly identified as threats; b) false alarm rate - the proportion of non-TIP bags incorrectly identified as threats; c) d′ - a signal detection measure of threat detection ability derived from hit and false alarm rates; and c - a signal detection measure of how sure a screener had to be to say a threat was present, also derived from hit and false alarm rates.

The performance-related validity of the CBT was examined with a correlation of content and image test scores with these TIP measures of performance. The focus was on both the initial and final content and image scores, for all screeners who had experienced at least 10 instances of TIP presentation.

## 3. RESULTS

Table 1 reports the total number of tests taken, average number of times trainees took each test unit, percentage of trainees that ultimately passed those units, the average score, and average score for the first time the unit was taken.

*Table 1.* Training module unit statistics.

| Unit | Number of tests taken | Average Number of Times | Percent Passed | Average Score | Average 1st Score |
|---|---|---|---|---|---|
| **1** | 872 | 1.32 | 75 | 76 | 82 |
| **2** | 936 | 1.46 | 68 | 72 | 77 |
| **3** | 975 | 1.56 | 64 | 70 | 78 |
| **4** | 853 | 1.40 | 72 | 73 | 78 |
| **5** | 896 | 1.49 | 67 | 71 | 74 |
| **6** | 1048 | 1.74 | 57 | 66 | 72 |
| **Content** | 1757 | 2.95 | 31 | 78 | 80 |
| **Image** | 1029 | 1.92 | 52 | 67 | 74 |

Three quantitative measures of item quality were calculated (error rates, item-to-test correlations, and FKR scores) and are reported below. Error rates were a measure of item difficulty. The item-to-test correlation (Phi) is a measure of item discrimination representing the relationship between performance on each individual question (pass/fail) and performance on the test unit containing that question (pass/fail). A high value of Phi means that performance on the test question correlates with performance on the unit test,

indicating that the question is predictive of overall performance. The FKR score provides a measure of item readability based on the number of words per sentence and the number of syllables per word. Table 2 provides the minimum, maximum, mean, and standard deviations for each of these measures.

*Table 2.* Error Rate, Phi, and FKR grade statistics.

| Measure | N | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|---|
| **Error rate** | 179 | 1.50 | 81.60 | 27.64 | 17.01 |
| **Phi** | 179 | -0.03 | 0.44 | 0.23 | 0.08 |
| **FKR grade** | 179 | 2.28 | 19.62 | 9.58 | 3.73 |

No significant relationships were found between error rate, Phi value, and FKR grade. The correlation coefficients between error rate and Phi value, error rate and FKR grade, and Phi value and FKR grade were all non-significant.

Error rates were examined in terms of grammaticality, semantic content, and wording indexes calculated for each question. None of these factors had significant effects on error rate. To further assess the quality of the test items, an analysis was performed on the hardest and easiest questions. The questions were ordered according to error rates and the ten questions with the lowest error rates and the ten questions with the highest error rates were examined. Two important patterns emerged and the first concerned the question response options. Half of the low error-rate questions contained response options that were unrealistic. For example, one question was - Clarifying your message, being understood, and respecting people's personal space are examples of: Interpersonal skills, Crisis, Conflicts, Potential violence.

Clearly, option 1 is the correct answer and options 2, 3, and 4 are nonsensical answers for this question. Options like these make the correct answer easy to detect by virtue of it being the only reasonable choice. None of the high error-rate questions contained unrealistic or absurd response options.

The second pattern concerns the role that general knowledge plays in answering the test questions. Seven of the ten low error-rate questions could be answered based on general knowledge and did not require the information presented in the CBT. Only one of the ten high error-rate questions could be answered this way (although this question required a sophisticated understanding of radiation, which was explained in the CBT). The implications of these two patterns will be addressed in the discussion section.

## 3.1 CBT Reliability

The CBT tests were constructed by taking a subset of items from a larger pool of items to make up individual tests. This has implications for measuring both internal validity and test-retest reliability. As this type of overall test design has some very useful properties

in terms of protecting test security, and hence preserving test validity, these problems are discussed at some length here.

Standard software to measure internal validity, such as the scale reliability analysis routines found in the Statistical Package for the Social Sciences, assume that every subject takes every question. If any question is missing, the whole case is omitted. This means that these routines cannot be used to look at the internal validity of an item data set like that produced by the CBT. This problem was avoided by measuring the inter-correlations of pairs of individual items. The algorithms that calculate coefficient alpha assume that differences in inter-item correlations are the result of random error. The same assumption is made here to calculate measures of test reliability equivalent to coefficient alpha by applying the Spearman-Brown formula to an average inter-item correlation

$$r_{TT} = \frac{k * r_{ij}}{1 + (k-1) * r_{ij}}$$

where $r_{ij}$ = the average inter-item correlation, and k = the number of items. As the current CBT tests were based upon instruction from six distinct units, a within-unit and between-unit average inter-item correlation was determined.

Three different item characteristics were used. The first was the consistency of responses to an item presented at two different times, the item test-retest coefficient. The last response to an item in one of the six unit tests was correlated with, if it existed, the response to the same item in the content test. This coefficient was pooled across all items and found it to be quite high (Phi = .42, $p<.01$).

Two item characteristics were also calculated based upon inter-item correlations. For a large sample of questions, two types of inter-item correlations were calculated for items in the final content test. The first was inter-item correlations of items that can appear within the same unit test. The second was correlations of items that only appear in different unit tests.

The average within unit inter-item correlation was 0.08, and the average between unit inter-item correlation was 0.025. This means that item reliability was very high using test-retest measures (the Phi = .42 above) and very low using these two uniformity measures. This reflects heterogeneous test content, but also is problematic for a test that samples from a pool of questions. The 50-item content test was a mixture of items from the same unit and items from different units. If given a second time, some of these items will be repeated. An estimate of reliability of the test based upon the between unit correlations and the Spearman-Brown formula would lead to an estimate of internal validity of 0.56. On the other hand, if the test was simply repeated without re-sampling items, the estimated test-retest reliability is 0.97. The number of items repeated from one presentation of the content test to another was calculated. Based upon this sample, 23.5 items were repeated. This is more than would be expected by random sampling.

## 3.2 Adverse Impact

Two main issues were pursued in the analysis of adverse impact. It was first determined whether there was differential success in training associated with race or gender. Then screeners' initial scores on tests were used to determine whether adverse impact would exist if testing were limited to a single administration.

The number of screeners in each racial and gender group who began the training, and the number who successfully completed the training, is given in Table 3.

*Table 3.* Training success rates for racial and gender groups.

| | Began Training | Completed Training | Percentage Successful |
|---|---|---|---|
| **Asian** | 15 | 5 | 33 |
| **Black** | 132 | 92 | 70 |
| **Hispanic** | 15 | 6 | 40 |
| **Other** | 9 | 7 | 78 |
| **White** | 32 | 24 | 75 |
| | | | |
| **Females** | 129 | 85 | 66 |
| **Males** | 79 | 54 | 68 |
| | | | |

Chi Square Analysis showed that there was a significant adverse impact associated with Asians and Hispanics relative to whites ($p<.01$), and no effect of gender ($p = .71$) on the success rate.

Another factor potentially affecting success rate is whether or not the screener is a native English speaker. There was not a statistically significant difference in CBT pass percentage between native English speakers and those speaking English as a second language. While 64% of Asians and Hispanics who described themselves as nonnative speakers failed to complete the CBT, 62% of those who said they were native speakers also failed to complete the CBT.

Interestingly, the examination of test scores provides a slightly different picture of adverse impact compared to the examination of success rates. The initial examination of test scores was restricted to the first time a screener took a unit test. An analysis of variance with unit, gender, and race as factors was performed. The interaction of gender and race [$F(4,1328) = 6.1, p<.01$], as well as the main effects of race [$F(4,1328) = 13.6$, $p<.01$] and unit number [$F(7,1352) = 5.1, p<.01$] were significant. Posthoc analyses showed whites scored higher than every group but Asians. Mean scores by race and gender are shown in Table 4. These suggest adverse impact in the SPI test, providing additional rational for the FAA validating a standardized SRT that does not adversely impact.

*Table 4.* Average test scores for racial/ethnic and gender groups.

| | Females | Males |
|---|---|---|
| Asian | 71 | 76 |
| African-American | 70 | 70 |
| Hispanic | 73 | 51 |
| White | 79 | 77 |
| Other | 62 | 70 |

The initial unit pass rates for the different groups were Asians 68%, blacks 59%, Hispanics 56%, others 43% and whites 73%. Thus, the different pass rates for the groups reflected the different mean scores [$\Psi^2$ (10) = 34.05, $p$<.01].

An analysis of covariance using language status as a co-variate still yielded significant effects of race, indicating that the scoring differences are not accounted for by differences in native language.

A second analysis of variance was performed restricted to first-time performance on the final CBT tests of content and images. This is a more restricted range of screeners as large numbers did not successfully negotiate units 1-6 to reach this point. Within this restricted group, there are no significant differences between racial groups. The mean unit scores are shown in Tables 5 and 6. Only the effect of unit number was significant [F(1,281) = 10.6, $p$<.01]. Image unit scores were lower than the content unit scores.

It was not possible to examine issues of differential validity because of the lack of a good demographic mixture for the TIP data. It was, however, possible to examine DIF.

*Table 5.* Racial and gender breakdown of content scores.

| **GENDER** | **RACE** | **N** | **Mean** |
|---|---|---|---|
| **Female** | **Asian** | 4 | 88 |
| | **Black** | 61 | 76 |
| | **Hispanic** | 8 | 72 |
| | **Other** | 5 | 72 |
| | **White** | 16 | 80 |
| | | | |
| **Male** | **Asian** | 8 | 77 |
| | **Black** | 43 | 73 |
| | **Hispanic** | 2 | 58 |
| | **Other** | 2 | 77 |
| | **White** | 12 | 81 |

A significant DIF difference was found for 35 of the 179 test questions, though a small number (5) of these questions were found to be biased for, rather than against, the non-white group. No strong determinants of bias in the questions could be identified. A

small difference in FKR in DIF questions (10.0) compared to non-DIF (9.5) questions was not significant.

*Table 6.* Racial and gender breakdown of image scores.

| Gender | Race | N | Mean |
|---|---|---|---|
| **Female** | **Asian** | 2 | 92 |
| | **Black** | 57 | 61 |
| | **Hispanic** | 6 | 70 |
| | **Other** | 5 | 45 |
| | **White** | 14 | 65 |
| | | | |
| **Male** | **Asian** | 5 | 59 |
| | **Black** | 36 | 62 |
| | **Hispanic** | 0 | |
| | **Other** | 2 | 59 |
| | **White** | 12 | 61 |
| | | | |

Below are two questions that exhibited DIF by way of illustration. Tables 7 and 8 show the pass rates of these questions for different ability and racial groups. The following question showed significant DIF: Who are the primary people assigned to the security checkpoint? Preboard screeners and a ground security coordinator, Preboard screeners and a Checkpoint Security Supervisor*, A Ground Security Coordinator and a Checkpoint Security Supervisor, A Law Enforcement Officer and a Ground Security Coordinator.

*Table 7.* Unit and item pass-rates for different racial groups for a single item.

| Unit | Item | White | Minority |
|---|---|---|---|
| PASS | Correct | 17 (89%) | 115 (82%) |
| | Wrong | 2 | 26 |
| | Total | 19 | 141 |
| | | | |
| FAIL | Correct | 20 (87%) | 105 (58%) |
| | Wrong | 3 | 76 |
| | Total | 23 | 181 |
| | | | |

While 17 of 19 (89%) whites who passed the test answered the item correctly, a slightly lower percentage, 115 of 141 (82%), of non-whites did so. The 20 of 23 (87%) whites who failed the test answered the question correctly. Only 105 of 181 (58%) non-whites did so. This is an example of DIF for two reasons. While the item discriminates between non-whites of high and low ability, it does not do so with whites. Additionally, whites in every ability group scored above non-whites on this item, a racial difference in performance exists when ability is equated.

A second example is - The efforts of the FAA and the ICAO to develop security regulations to screen passengers can be called? Security Program*, Security Control, Security Survey, Security Restrictions.

*Table 8.* Unit and item pass-rates for different racial groups for a single item.

| Unit | Item | White | Minority |
|---|---|---|---|
| Pass | Correct | 25 (89%) | 78 (66%) |
| | Wrong | 3 | 41 |
| | Total | 28 | 119 |
| | | | |
| Fail | Correct | 11 (73%) | 49 (49%) |
| | Wrong | 4 | 50 |
| | Total | 15 | 99 |
| | | | |

Similar to the first example, 25 of 28 (89%) whites who passed the test answered the item correctly while a lower percentage, 78 of 119 (66%), of minorities did so. Thus, 11 of 15 (73%) whites failed the test, but answered the question correctly, while only 49 of 99 (49%) minorities did so. As before, whites in every ability group scored above non-whites on this item. While a racial difference in performance exists for this item when ability is equated, it is important to note that the content of this question does not reflect any knowledge or skill that is required for performing any screener duties. It is interesting to note that of the 35 test items that showed adverse impact, 14 (40%) were judged to be irrelevant to on-the-job performance. The presence of non-relevant questions should be considered when evaluating the efficacy of screener-competence tests and in determining which questions should be removed, especially when they are found to have race-based differential functioning.

## 3.3 Criterion Validity

TIP data were available for a group of the screeners, primarily from ATL. For each screener, hit rate, false alarm rate, d′, and c were calculated. The correlation of these variables with performance on the final content and image tests was examined. These correlations are shown in Table 9 and performance for the first time the test was taken and performance for the final time the test was taken were used. All these measures are subject to some range restriction because individuals who failed CBT did not go on to become screeners and have no TIP data available. The sample was further restricted to individuals who had been exposed to at least 10 threats with the TIP system in order to reduce unreliability in the TIP measures. With these restrictions, the sample of screeners was 76.

*Table 9.* Correlations between TIP performance measures and scores on final content and image tests.

| | First Content | First Image | Last Content | Last Image |
|---|---|---|---|---|
| Hit rate | .24* | .24* | .23* | .09 |
| FA rate | .18 | -.14 | .04 | -.02 |
| d′ | .12 | .28* | .18 | .08 |
| c | .23* | .07 | .12 | .03 |

Hit rate was significantly correlated with the initial content, image test scores, and final content test score. There were no significant correlations with false alarm rate, but d′ was correlated with the initial image test and c was correlated with the initial content test.

A potential source of attenuation of the validity coefficient is unreliability of the criterion measure. The reliability of the TIP performance measures used was estimated in the following way. Data were restricted to months with at least four TIPs presented to a screener and then a monthly hit rate was calculated. Two independent estimates of screeners' hit rates were computed by using the even and the odd months. The average number of even months for a screener was 2.07, and the average number of odd months for a screener was 2.89. The two independent estimates were correlated and the split-half reliability was calculated using the Spearman–Brown formula. The split-half reliability of the average individual screener's TIP estimates was 0.88. Thus, the CBT-TIP correlations were not significantly affected by unreliability of the TIP measures.

### 3.4 Image Items

The image test used a weighted scoring algorithm which placed the greatest weight on correctly identifying the items from the FAA test item category. As a result, the most dramatic contrast between image tests that were passed and failed was the percentage of FAA test items identified. The mean number of items correctly identified for each class is reported in Table 10. A repeated measures analysis of variance was performed with item type as a within subject factor and pass or fail the test as a between subject factor. The effects of item type [$F(4,213) = 57.4$, $p < .001$], pass/fail performance [$F(1,213) = 99.3$, $p < .001$], and the interaction of item type and pass/fail performance [$F(4,213) = 16.1$, $p < .001$] were all significant.

*Table 10.* Mean number of FAA test items correctly identified in the SPI test prior to on-the-job training.

| | **FAA Articles** | **Guns** | **Innocent** | **Knives** | **Opaques** |
|---|---|---|---|---|---|
| **Pass Test** | .82 | .97 | .87 | .87 | .86 |
| **Fail Test** | .49 | .86 | .70 | .73 | .73 |
| **Overall** | .65 | .92 | .78 | .79 | .79 |

Even including failed image tests, scores were quite respectable for every class of definite and possible threat, except for FAA articles in tests where the test is not passed. The false alarm rates were high and would not be practical in a field situation. The d′ and c were calculated for each test and for the TIP data. These values are presented in the Table 11.

*Table 11.* The d′ and c for each test and TIP.

| | **Image / Overall** | **Image / Fail** | **Image / Pass** | **TIP** |
|---|---|---|---|---|
| **d′** | 1.76 | 1.07 | 2.50 | 2.74 |
| **c** | -.05 | -.06 | -.04 | -1.61 |

There are substantial differences in the two data sets when overall test performance is compared with TIP. When the data set is restricted to the final tests taken by those who passed the CBT, the differences in d′ largely disappear. In this situation, the most dramatic difference between the TIP data and the image set data was in the criterion c. Working screeners using TIP are less likely to call an innocent bag a threat than are trainees taking the CBT image test.

## 4. DISCUSSION

### 4.1 Question Quality

The overall findings indicated no effects of item readability (i.e., FKR grade level, grammar, semantic clarity, or wording) on overall performance. Furthermore, there were no significant interactions with item-to-test correlations. While items varied in terms of how well they predicted overall performance, there were no systematic patterns found that could be used to identify *a priori* which questions were good test items and which were not.

When the content of the questions was examined, however, several interesting and important patterns were found. The results indicated two groups of questions that distinguished the low error-rate questions from high error-rate questions. Low error-rate questions had a larger number of items with poor response options (i.e., items in which the incorrect options were nonsensical, making them very easy to eliminate as possible responses). If a large percentage of the questions contain poor response options, test performance may reflect the subject's ability to rule out absurd choices and not the mastery of the subject matter being taught. This should be an important consideration when designing any new test battery and when assessing the efficacy of extant tests.

A second pattern indicated that many low error-rate questions could be answered based on one's general knowledge. For these questions, correct responses do not require acquisition nor retention of material presented in the CBT. Consequently, they should not be considered as accurate measures of learning. The presence of both types of questions (those that rely on general knowledge and those that require the CBT) in the

CBT exam may undermine the efficacy of the exam. To evaluate how well the CBT exam measures the information taught in the CBT, the impact of general knowledge questions must be assessed. This is best done by testing subjects naïve to the CBT' subject matter and comparing their data to that of subjects who have completed the CBT. The degree to which there is no difference in test scores between these two groups would indicate that the CBT test does not measure mastery of the information presented in the CBT. More importantly in this situation, good performance in the CBT test would support the argument that a significant proportion of the test items in the CBT test required a high level of general knowledge and did not reflect mastery of the content information present in the CBT.

## 4.2 Test Reliability

It was not possible to examine reliability of the image tests with the data available. The CBT content tests are tests in which items are drawn from a large pool of items. If the test is repeated, only a small proportion of items will be repeated. Differences in test-retest reliability and internal validity will be determined by the differences in item reliability (test–retest reliability for a single item) and inter-item correlations, as well as, the proportion of items that are repeated on a second testing. The most conservative estimate of reliability for a multiple choice test structured like this would be to use the average inter-item correlation and construct estimates of reliability for tests of different length using the Spearman-Brown formula. (Nunnally, 1978).

Given the empirically determined within-unit, inter-item correlation of 0.08 in this sample, a 10 item test with a reliability of 0.47, is rather low. This figure does not necessarily make the CBT unit tests unacceptable because they can be taken repeatedly. Because a comprehensive test will consist of heterogeneous content, test reliability is most accurately measured if an effort is made to analyze how heterogeneous content contributes to unreliability as measured by coefficient alpha. The best approach to measuring reliability, with multiple choice tests with heterogeneous content and item sampling from a larger pool, is to divide the test into sub-tests with more homogenous content and examine the inter-item correlations within the sub-tests. For a battery composed of component sub-tests, the overall reliability ($r_{tt}$) can be calculated based upon the reliability ($r_{NN}$). of the components (Nunnally, 1978). In the present case, the individual items are the components. The specific formula is shown below.

$$r_{TT} = 1 - \left( \frac{\sum \sigma_i^2 - \sum r_{NN} * \sigma_i^2}{\sigma_s^2} \right)$$

where, $\sigma_I$ is the standard deviation of item i and $\sigma_s$ is the standard deviation of sum of all items.

This formula can be adapted for the analysis of this type of multiple choice test. Specifically, the calculation can be based upon the average values which will convert the formula to the following:

$$r_{TT} = 1 - \left( \frac{(1 - r_{NN}) * \sigma_i^2}{\sigma_s^2} \right)$$

Even with this approach, the CBT in this case is characterized by low test reliability. This is mitigated by the fact that screeners can take tests repeatedly, but the lesson for item development in similar tests is that attending to inter-item correlations in test development is very important.

## 4.3 Adverse Impact

Adverse impact of the CBT was found for Asians and Hispanics with CBT completion rates. Adverse impact was found with blacks and Hispanics when initial CBT test scores were examined, but no adverse impact was found on first time content and image test scores. The implications of these results are perhaps clearest with the black group. Their average test scores on first taking any unit test were lower than those for whites. However, as a group they persist in the CBT, repeating unit tests rather than quitting. As a result, they finish the CBT at the same rate as whites do and showed no differences in performance by the time the content and image tests are taken.

The implication for multiple choice test development is that adverse impact may be associated not only with test items, but also with scoring protocol (e.g., whether individuals are allowed to repeat the tests, etc.), which can contribute to or mitigate adverse impact.

## 4.4 Test validity

Both the initial scores on the CBT content test and the initial scores on the CBT image test were found to be correlated with TIP hit rate. Additionally, image test scores were correlated with d′, while content test scores were correlated with c. On the other hand, the content test has low reliability (not able to estimate reliability of the image test). This implies that if the reliability of a set of multiple choice items could be increased substantially from what was found with the CBT, a significantly stronger relationship should have been found with the criterion variable.

## 4.5 Image Items

It was not possible to perform the same types of analyses with image items done with the smaller set of content items. However, some important things were learned about the image items and the image tests. The scoring algorithm of this test gave great weight to the category of FAA test articles which included all IEDs that were presented. From the differences in performance noted for these articles, and for other threats and innocent bags, overall performance level on an X-ray image test will be strongly influenced by the number of IEDs that are included in the test set. If a reliable image test is desired, then images should be drawn randomly from a larger set, and the proportions from each category of image be fixed for any administration of the test.

All adverse impact found was associated with the content tests. No adverse impact was found with the image tests. Since these tests are not taken until all other training and testing was completed, it could not be determined whether the absence of adverse impact was a function of differential success rates or not. The image test correlated with the most important measure of TIP performance, d', in the field. More than that, this particular image test produced results measured in d' which were very comparable with TIP performance in the field even though the response criterion seen in the test and in the field were very different.

## 5. FUTURE APPLICATIONS

At least three organizations are currently developing CBT systems for checkpoint screeners (Air Transport Association, Frontline International, and ICTS). These efforts can be generally supported by the FAA's Aviation Security Human Factors Program's extensive guidance on the training development process for aviation security screeners (Fobes & Neiderman, 1997). The new CBT systems being developed may very well have built-in mastery testing similar to the testing used in the SPI training (Neiderman & Fobes, 1998). The present report on validating CBT mastery tests found that testing can be a reliable and valid measure of learning if properly developed. That is, care must be exercised, during the development of test items, to guarantee the test will have relevance to job requirements and efforts must be made to obtain reliable and valid measures. Since test items can exhibit DIF without obvious cause, adverse impact should be tested for and the role of scoring protocols in promoting and mitigating adverse impact explored. Correlations found with TIP performance indicate that a multiple choice content and image test can be an effective test of screener readiness to progress to on-the-job training.

This report provides additional guidance on the training development process. However, the FAA should not rely upon training developers to determine the final exam by which a screener candidate is judged to have successfully completed knowledge acquisition during the initial training and to be ready for the on-the-job phase of acquiring skills and abilities. The FAA's Aviation Security Human Factors Program is developing a SRT consisting of multiple choice questions about the screeners' job and X-ray image interpretation test. The material in the present evaluation will additionally provide information useful to SRT development as well as for the developer's evaluation of built-in testing in their new CBT system.

## 6. REFERENCES

Camilli,G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Sage, Thousand Oaks, CA.

Fobes, J. L. & Neiderman, E. (1997). *The Training Development Process for Aviation Screeners*. Technical Report DOT/FAA/AR-97/46, DOT/FAA Technical Center, Atlantic City International Airport, NJ.

Fobes, J. L. & Neiderman, E. (1998). *Screener Readiness Testing Preliminary Item Analyses*. Technical Report DOT/FAA/AR-98/39, DOT/FAA Technical Center, Atlantic City International Airport, NJ.

Hulin, C., Drasgow, T., & Parsons, C. (1983*). Item Response Theory: Application to Psychological Measurement.* Dow-Jones-Irwin: Homewood, IL

Nunnally, Jum C. (1978). *Psychometric Theory*, McGraw Hill N.Y., N.Y.

*Uniform Guidelines on Employee Selection Procedures* (1978); 43 FR 38295.